Computers and Mathematics with Applications 58 (2009) 1930–1935

Contents lists available at ScienceDirect

Computers and Mathematics with Applications

journal homepage: www.elsevier.com/locate/camwa

Classic group testing with cost for grouping and testing

Danny W. Turner [a], James D. Stamey [b,*], Dean M. Young [b]

[a] Department of Mathematics, Winthrop University, Rock Hill, SC 29733, USA
[b] Department of Statistical Science, Baylor University, Waco, TX 76798-7140, USA

ARTICLE INFO

Article history:
Received 2 November 2007
Received in revised form 18 February 2009
Accepted 18 March 2009

Keywords:
Blood testing
Cost functions
Expected costs
Nonlinear integer programming

ABSTRACT

We generalize the classical group testing problem to incorporate costs associated with pooling and inspection, both of which are significant factors in actual applications. We formulate the expected cost model as a nonlinear integer programming problem, prove several propositions and a theorem concerning when pooling is more efficient than individual testing, and determine the optimal group size such that the expected cost is minimized.

1. Introduction

The classic blood testing problem was introduced by Dorfman [1] and involved the inspection of members of a large population for some disease by using a blood test which registers positive if a sample contains blood from a person having the disease and registers negative otherwise. Since the inspection process may be expensive and/or time consuming, a two-stage strategy is suggested. In the first stage, the blood samples of groups of k persons are pooled, and the pooled samples are tested. If a pooled sample registers negative, the group is cleared; otherwise, the inspection procedure moves to stage two, and each person in the positive group is tested. Of course, this two-stage group screening procedure is potentially useful whenever grouping and testing are feasible. For example, batteries might be tested in groups to detect dead cells, or an expensive test can be run on pooled samples to detect the presence of a certain rare contaminant in well water samples.

This scenario involving group testing has received much attention in the literature. Material dealing specifically with Dorfman's basic problem can be found in a number of references, including, for example, [2–11]. Dorfman's group testing problem has been generalized in various ways and adapted for a wide variety of applications. From a statistical estimation perspective, this problem has been considered by Kennedy [12], Vansteelandt, Goetghebeur, and Verstraeten [13], and McCann and Tebbs [14]. While estimation is not our goal here, in applications where the parameters are unknown, incorporating techniques from the above articles would be of interest.

Our generalization of the original problem incorporates costs associated with pooling and inspection, both of which would be factors in a realistic application. We set up the expected cost model as a nonlinear integer programming problem and determine the optimal group size such that the expected cost is minimized.

2. The model

Suppose a test is to be administered to a large number of units to inspect for some specified characteristic. Assume that the population size N is known, that the probability p of a randomly selected unit having a positive test, i.e., having the

* Corresponding author.
E-mail address: James_Stamey@baylor.edu (J.D. Stamey).

0898-1221/$ – see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.camwa.2009.03.119

characteristic, is the same for all N units, that these N units are stochastically independent, and that the costs for pooling and testing are known. Also, we assume that any "leftover" group of size ($N \mod k$) is automatically pooled. Our goal is to determine situations where pooled testing of groups of size k and any pooled leftovers reduce the total expected cost below that of simply testing each unit separately. The testing process is as follows.

If a pooled sample yields a negative result, this one test suffices to clear this group (of size k or ($N \mod k$)). If the test yields a positive result for a pooled sample, then each unit in this group will be tested individually. Our generalized group testing problem is to find the value of k that minimizes the expected total cost of the testing procedure. Optimal values of k may not be unique.

We use the following definitions:

$N \equiv$ population size;

$k \equiv$ common size of each complete group, $1 \le k \le N$;

$C_1(k) \equiv$ cost per group of pooling with group size k;

$C_2(k) \equiv$ cost of testing a pooled sample for a group of size k;

$C_3 \equiv$ cost of testing one unit (constant);

$G(N, k) \equiv$ integer part of $N/k =$ number of complete groups of size k that can be formed;

$p \equiv$ probability that a randomly selected unit has the specified characteristic;

$q \equiv 1 - p$.

Also, we assume that $C_i(k) \ge 0$ for all $k > 0, i = 1, 2, 3$, that $0 < p < 1$, and that the test itself is infallible for both an individual unit and a pooled sample. The costs associated with pooling and testing the pooled samples are, for a given N, functions of k so that the total costs associated with these two operations are deterministic. Observe, for example, that the units of cost could be time rather than dollars.

Note that $C_1(1) = C_2(1) = 0$ because no pooling costs and no pooled sample testing costs exist when the group size is $k = 1$. Therefore, costs related to the pooled groups (including the leftover group, if there is one) are:

$$\text{Pooling Cost} \equiv \begin{cases} G(N,k)C_1(k) + C_1(N \bmod k), & \text{if } N/k > G(N,k) \\ G(N,k)C_1(k) & \text{if } N/k = G(N,k) \end{cases}$$

and

$$\text{Testing Cost} \equiv \begin{cases} G(N,k)C_2(k) + C_2(N \bmod k), & \text{if } N/k > G(N,k) \\ G(N,k)C_2(k), & \text{if } N/k = G(N,k). \end{cases}$$

If we let $L(N, k) \equiv N \mod k$ and $C_1(0) = C_2(0) \equiv 0$, then the above cost formulas can be rewritten in the more compact form

$$\text{Cost}_i = G(N,k)C_i(k) + C_i(L(N,k)),$$

where $i = 1$ denotes the pooling cost and $i = 2$ gives the testing cost.

Next, consider costs related to individual testing. The number of positive results obtained for the $G(N, k)$ pooled samples, each of size k, has a binomial distribution with parameters $G(N, k)$ and $1 - q^k$. Thus $G(N, k)(1 - q^k)$ is the expected number of positive pooled samples that will require testing of individual group members. Hence, the expected individual testing cost for $G(N, k)$ whenever $k > 1$ is

$$E[\text{Cost}_{G(N,k)}] = kG(N,k)(1 - q^k)C_3.$$

Similarly, the expected cost of individual testing for the leftover group is

$$E[\text{Cost}_{L(N,k)}] = (N - kG(N,k))(1 - q^{N-kG(N,k)})C_3,$$

provided that $k > 1$ and $N - kG(N, k) > 1$. If $N - kG(N, k) = 1$, then the leftover group has a single unit that must be inspected with a cost of C_3. Let

$$Q(q; N, k) \equiv \begin{cases} q^{N-kG(N,k)}, & \text{if } N - kG(N,k) \ne 1 \\ 0, & \text{if } N - kG(N,k) = 1. \end{cases}$$

The expected individual testing cost for the leftover group can now be written as $E[\text{Cost}_{L(N,k)}] = (N - kG(N, k))(1 - Q(q; N, k))C_3, k > 1$. When we combine our two costs, we determine that $E[\text{Total Cost}_N] = G(N, k)\{C_1(k) + C_2(k)\} + C_1(L(N, k)) + C_2(L(N, k)) + \{kG(N, k)(1 - q^k) + (N - kG(N, k))(1 - Q(q; N, k))\}C_3$ for $k > 1$. The relative expected cost (i.e., average cost per unit) is the above expected total cost divided by N for $k > 1$. If $k = 1$, then no pooling cost exists and the total individual testing cost is NC_3 and, thus, the associated relative cost is obviously C_3. Now let $C(k) = C_1(k) + C_2(k)$ and $a = C_3$. Then, the relative expected cost, $f(k) \equiv E[\text{Total Cost}_N]/N$, is

$$f(k) = \begin{cases} a, & \text{if } k = 1 \\ (1/N)G(N,k)C(k) + (1/N)C(L(N,k)) + a\{(k/N)G(N,k)(1 - q^k) \\ \quad + (1 - Q(q; N, k))(1 - (k/N)G(N,k))\}, & \text{if } 1 < k \le N. \end{cases} \tag{1}$$

Having established a model for relative expected cost, we next address the problem of determining the optimal group size k so that the relative expected cost is minimized. For $k > 1, f(k)$ in (1) can be rewritten as

$$f(k) = a + (1/N)G(N,k)C(k) + (1/N)C(N - kG(N,k)) - (a/N)\{kG(N,k)q^k + Q(q;N,k)(N - kG(N,k))\}. \tag{2}$$

If we let

$$F(k) \equiv \begin{cases} 0, & \text{if } k = 1 \\ G(N,k)C(k) + C(N - kG(N,k)) - a\{kG(N,k)q^k + Q(q;N,k)(N - kG(N,k))\}, & \text{if } 1 < k \le N, \end{cases} \tag{3}$$

it follows from (1)–(3) that

$$f(k) = a + (1/N)F(k). \tag{4}$$

Because q and N are positive constants, one can see from (4) that the optimization problem that we wish to solve is the following nonlinear integer programming model:

Minimize $F(k)$ given in (3), subject to the constraint that k is an integer such that $1 \le k \le N$. (5)

3. Analysis

We next develop several inequalities concerning the optimal group size in four propositions and a theorem. To begin with, note that because $kG(N,k)$ is positive and $N - kG(N,k)$ is nonnegative, we have that

$$A_1(k) = kG(N,k)q^k + Q(q;N,k)(N - kG(N,k)) > 0$$

and

$$A_2(k) = G(N,k)C(k) + C(N - kG(N,k)) \ge 0.$$

Proposition 1. *The optimal group size is $k^* > 1$ if and only if there exists an integer $k > 1$ such that $F(k) < 0$.*

Proof. This follows immediately from (4) because the relative expected cost is a when $k = 1$. □

Proposition 2. *Suppose that the cost function $C(k)$ is nondecreasing for $2 \le k \le N$. If there is no integer $k > 1$ such that $aA_1(k) > C(2)$, then the optimal group size is $k^* = 1$. Moreover, if there exists an integer $k' > 1$ such that $aA_1(k') > (1 + N/2)C(N)$, then the optimal group size is $k^* > 1$.*

Proof. First, observe that $C(k) \ge C(N - kG(N,k))$ because $k > N - kG(N,k)$. Also, we have that $1 \le G(N,k) \le G(N,2)$ because $2 \le k \le N$. Therefore, $G(N,k)C(k) \le A_2(k) \le (1 + G(N,k))C(k)$. Moreover, $C(2) \le C(k) \le G(N,k)C(k)$ and $(1 + G(N,k))C(k) \le (1 + N/2)C(N)$ yield that $C(2) \le A_2(k) \le (1 + N/2)C(N)$, from which we obtain $C(2) - aA_1(k) \le F(k) = A_2(k) - aA_1(k) \le (1 + N/2)C(N) - aA_1(k)$. The result follows from the last sequence of inequalities. □

Proposition 3. *Suppose $C(k)$ is nonincreasing for $2 \le k \le N$. If $aA_1(k) > C(N)$ fails to hold for all $k > 1$, then the optimal group size is $k^* = 1$, and if there exists an integer $k' > 1$ for which $aA_1(k') > (1 + N/2)C(2)$, then the optimal group size is $k^* > 1$.*

Proof. Note from the definition of $A_2(k)$ that $G(N,k)C(k) \le A_2(k)$. When $k \ge 2$, then $C(N) \le C(k) \le G(N,k)C(k)$ and $C(N - kG(N,k)) \le C(2)$, from which we have that $C(N) \le A_2(k) \le (1 + N/2)C(2)$. This fact implies that $C(N) - aA_1(k) \le F(k) = A_2(k) - aA_1(k) \le (1 + N/2)C(2) - aA_1(k)$. The proposition follows immediately from the last set of inequalities. □

Proposition 4. *Suppose $m \le C(k) \le M$ for $2 \le k \le N$, where m and M are positive constants. If there is no integer $k > 1$ such that $aA_1(k) > m$, then the optimal group size is $k^* = 1$, and if there does exist an integer $k' > 1$ such that $aA_1(k') > (1 + N/2)M$, then the optimal group size is $k^* > 1$.*

Proof. The result follows by noting that $m - aA_1(k) \le F(k) = A_2(k) - aA_1(k) \le (1 + N/2)M - aA_1(k)$. □

Theorem. *Suppose the cost function satisfies $C(k) = c$ for $2 \le k \le N$, where c is a positive real number. If $N - kG(N,k) > 1$ and $(a/c) < 1/N$, then the optimal group size is $k^* = 1$. If $N - kG(N,k) \le 1$ and $(a/c) > (1 + N/2)/((N - 1)q^N)$, then the optimal group size is $k^* > 1$.*

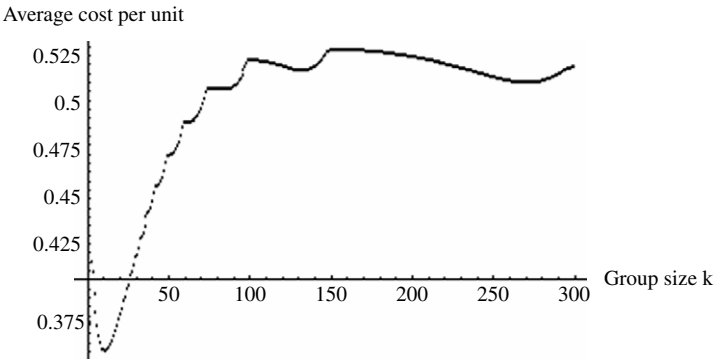

Fig. 1. Relative expected cost as function of group size for $p = 0.03$.

Proof. Let $b \equiv a/c$ and let

$$\delta \equiv \begin{cases} 1, & \text{if } N - kG(N,k) > 1 \\ 0, & \text{otherwise.} \end{cases}$$

From (3) with $k > 1$, we have that

$$F(k) = c\{\delta + G(N,k) - b(kG(N,k)q^k + Q(q;N,k)(N - kG(N,k)))\}. \tag{6}$$

Observe that if $k > 1$, then $G(N,k)c - a\{kG(N,k)q^k + Q(q;N,k)(N - kG(N,k))\} \le F(k) \le (1 + G(N,k))c - a\{kG(N,k)q^k + Q(q;N,k)(N - kG(N,k))\}$. Noting that $(N - kG(N,k)) < k$, we see that if $N - kG(N,k) > 1$, then $Nq^k \le kG(N,k)q^k + Q(q;N,k)(N - kG(N,k)) \le NQ(q;N,k)$, which implies that $G(N,k)c - aNQ(q;N,k) \le F(k) \le (1 + G(N,k))c - aNq^k$. Therefore, $c - aN \le F(k) \le (1 + N/2)c - aNq^N$ and so $c(1 - bN) \le F(k) \le c(1 + N/2 - bNq^N)$. Thus, if $N - k[N/k] \le 1$, then $c(1 - bN) \le F(k) \le c(1 + N/2 - bNq^N)$ because $N > 1$. Therefore, if $b > (1 + N/2)/((N-1)q^N)$, then $F(k) < 0$ and we must have that $k^* > 1$. □

For many applications N is large. The above inequality shows us that if the cost of testing one unit is smaller than the sum of the costs of pooling per group and testing for each pooled sample, i.e., $b = a/c$ is very small, then the best procedure is to inspect units individually. Similarly, $1/(2q^N)$ is large when N is large, and thus when b is large, the two-stage strategy should be invoked.

The above results provide interesting theory, but from an applied point of view one can find the optimal group size k^* and the corresponding minimum relative expected cost $f(k^*)$ for various choices of p, N, and $C(\cdot)$ relatively easily. To do so, we compute the sequence $f(1), f(2), f(3), \ldots, f(N)$ and then determine the optimal group sizes as those values of k that correspond to the smallest $f(k)$ in this sequence. This approach is illustrated in Section 4.

4. Solving for the optimal group sizes

Solving the integer optimization problem in (5) is typically routine when one uses a computer algebra system such as *Mathematica*, Wolfram [15]. The ability to do exact computations is crucial in identifying multiple optimal solutions. In this section we illustrate how to use *Mathematica* to solve our group testing expected cost minimization problem. Portions of a *Mathematica* notebook used for our examples are shown in the Appendix. This *Mathematica* code can easily be modified to solve a wide variety of problem scenarios. Optimal solutions and graphical information concerning the behavior of the relative expected cost function for various problem configurations requires only moderate expertise in *Mathematica*.

The first example we illustrate involves $N = 300$ units where units have probability $p = 0.03$ of testing positive (i.e., units have a 3% chance of having the characteristic in question). The cost (\$) of testing one unit is $a = C_3 = 0.50$, and the cost function (\$) for pooling and testing a group of size k is $C(k) = \ln(k)$, $k > 1$. The output, plotted in Fig. 1, indicates that groups of size 12 are optimal and the resulting minimized average cost per unit is \$0.360154, resulting in an average savings of $300(0.50 - 0.360154) = \$41.95$ for each shipment of 300 units that needs to be screened. The plot of relative expected cost versus group size visually confirms our findings while showing the oscillatory behavior of the average cost per unit as a function of k for the given values of $N = 300$ and $q = 0.97$. Note that lower case n is used in the code to avoid confusion with *Mathematica's* use of the symbol N. We note that if a cost function of $C(k) = 1$, $k > 1$ were used instead of $C(k) = \ln(k)$, $k > 1$ the optimal group sizes would be 16 instead of 12, whereas if $C(k) = \sqrt{k}$, $k > 1$ the optimal group size would be 19. Thus there is some sensitivity to choice of cost function and this issue should be carefully considered.

Interestingly, we note that if p is changed from 0.03 to $p = 0.01$ in the above example, the optimal group increases to $k^* = 25$, the minimum relative expected cost becomes 0.239844 with a corresponding average savings for screening 300 units of $300(0.50 - 0.239844) = \$78.05$. This example illustrates the potential sensitivity of the optimal solution to the model parameters.

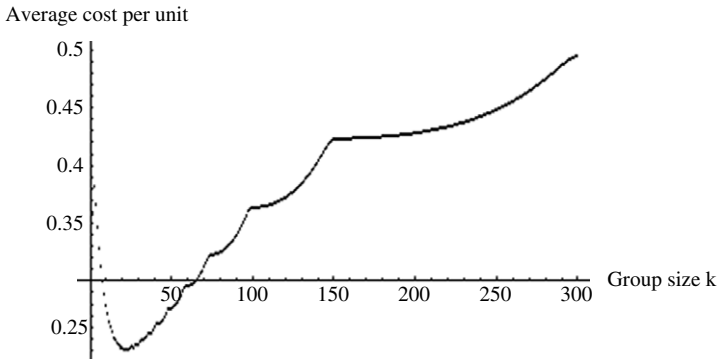

Fig. 2. Relative expected cost as function of group size for $p = 0.01$.

Continuing the above scenario, suppose the cost increases from \$.50 to \$.75 for testing a unit. For the case of $p = 0.03$, we determine that the solution is $k^* = 2$ and is $k^* = 17$ for $p = 0.01$ (Fig. 2).

Another nice example is to use $C(k) = 1$, $k > 1$, $a = 1.0$, $N = 300$, and $q = (1/3)^{1/3}$. For this model, two optimal group sizes (either $k^* = 1$ or $k^* = 3$) exist, and the minimum relative expected cost is \$1.00 for either choice. Example code is provided in the Appendix, and one can obtain the *Mathematica* programs for this example from the authors.

5. Summary

In this paper we have generalized the classical group testing problem to incorporate costs associated with pooling and inspection. These costs in terms of money and/or time are clearly present in actual group testing applications and should be accounted for in determining the optimal group testing group size solution. Assuming an infallible test, we formulate the expected cost model as a nonlinear integer programming problem and prove several propositions and a theorem concerning when pooling is more efficient than individual testing. Last, we use the software *Mathematica* to determine the optimal group size that minimizes the expected cost function, which incorporates pooling and inspection costs. We present several example scenarios and their solutions.

Many screening tests, both industrial and medical, are not infallible. For instance, Dendukuri and Joseph [16] discuss two fallible tests for *Strongyloides* infection. The procedure we have proposed is best suited for industrial applications where misclassification rates are small and consequences of misclassification are minimal. In future research we will consider incorporating misclassification error into the procedure.

Acknowledgement

The authors are grateful to a referee for the insightful comments that improved the paper.

Appendix

In this appendix we provide the *Mathematica* code used in the examples in Section 4.

- The first step is to set up the basic formulas. Use exact numbers for input parameters/constants to obtain exact results.

```
Clear[costmodelsetup, Cost, a, n, k, q, y, Q, G, m, L, F]
costmodelsetup[aval_, Cost_] :=
  ( Cost[0] := 0; Cost[1] := 0 ; a = aval;
   Q[q_, 0] := 1;
 Q[q_, 1] := 0;
 Q[q_, y_] := q^y;
 G[n_, k_] := Floor[n/k];
 m[n_, k_] := k * G[n, k];
 L[n_, k_] := n - m[n, k];
 F[q_, n_, 1] := 0;
 F[q_, n_, k_] := (G[n, k] * Cost[k] + Cost[L[n, k]] - a * (m[n, k] * q^k + L[n, k] * Q[q, L[n, k]]));
 RelExpCost[q_, n_, k_] := a + (1/n) * F[q, n, k]);
```

- Solve the nonlinear integer optimization problem to minimize the relative expected cost. Lower case n is used for the population size to avoid conflict with the built-in N function.

```
Clear[optimize, costs, mincost]
optimize[p_, n_] :=
 Module[{p1, n1}, p1 = p; n1 = n; q1 = 1 - p1;
  Print["p =  ", p1, "     q =  ", q1, "     n =  ", n1];
Print["Cost of testing one unit is a = ", a];
Print["For k>1, Cost function C(k) is  ", Cost[k]];
costs = Table[RelExpCost[q1, n1, k], {k, 1, n1}];
mincost = Min[costs];
Print["Minimum Relative expected cost {exact, approximate}:  ",
   {mincost, N[mincost, 10]}];
Do[If[mincost == costs[[k]], Print["Optimal group size(s):   ", k]], {k, 1, n}];
Print["Average resource savings = ", N[n (a - mincost), 10]];
  ListPlot[costs, PlotStyle → {PointSize[.009]},
   AxesLabel → {"Group size k", "Average cost per unit"}];]
```

- Enter the cost function (combined cost of pooling and testing a group of size k) and the value of a (cost of testing one individual). The combined cost can be decomposed into the component functions if desired.
- Define the specific cost formulas and apply our setup command.

```
Cost[k_] := Log[k]
costmodelsetup[1/2, Cost];
```

- Optimize for specific values of:
  p (probability of a positive test for a single unit) and
  n (population size).

```
optimize[3/100, 300]
```

OUTPUT:

```
p =  3/100     q =  97/100     n =  300

Cost of testing one unit is a =  1/2
Minimum Relative expected cost {exact, approximate}:
 {1/2 + 1/300 (-208152708298631400088512 3/2000000000000000000000 + 25 Log[12]), 0.3601543737}

Optimal group size(s):   12

Average resource savings = 41.95368790
```

References

[1] R. Dorfman, The detection of defective members of large populations, Ann. Math. Statist. 14 (1948) 436–440.
[2] W. Feller, An Introduction to Probability Theory and its Applications, vol. I, 3rd ed., John Wiley, New York, 1968.
[3] S.S. Wilks, Mathematical Statistics, John Wiley, New York, 1962.
[4] J.B. Pomeranz, On pooled testing, Math. Sci. 1 (1976) 99–106.
[5] S. Samuels, The exact solution to the two-stage group-testing problem, Technometrics 20 (1978) 497–500.
[6] D.W. Turner, F.E. Tidmore, D.M. Young, A calculus based approach to the blood testing problem, SIAM Rev. 30 (1988) 119–122.
[7] Q. Huang, N.L. Johnson, S. Kotz, Modified Dorfman–Sterrett screening (group testing) procedures and the effects of faulty inspection, Comm. Statist. Theory Methods 18 (1988) 1485–1495.
[8] N.L. Johnson, S. Kotz, R.N. Rodriguez, Dorfman–Sterrett screening (group testing) schemes and the effects of faulty inspection, Comm. Statist. Theory Methods 18 (1989) 1469–1484.
[9] N.L. Johnson, S. Kotz, Q. Wang, Randomized-sequential group testing procedures, Sequential Anal. 9 (1990) 219–242.
[10] X.Z. Wu, B.J. Zhao, Optimal sampling of hierarchical screening with inspection errors, Comm. Statist. Theory Methods 23 (1988) 803–814.
[11] V.A. Lancaster, S. Keller-McNulty, A review of composite sampling methods, J. Amer. Statist. Assoc. 93 (1998) 1216–1230.
[12] N. Kennedy, Multistage group testing procedure (group screening), Comm. Statist. Simulation Comput. 34 (2004) 621–637.
[13] S. Vansteelandt, J. Goetghebeur, T. Verstraeten, Regression models for disease prevalence with diagnostic rates on pools of serum samples, Biometrics 56 (2000) 1126–1133.
[14] M. McCann, J. Tebbs, Pairwise comparisons for proportions estimated by pooled testing, J. Statist. Plann. Inference 137 (2007) 1278–1290.
[15] S. Wolfram, The Mathematica Book, 4th ed., Cambridge University Press, 1999.
[16] N. Dendukuri, L. Joseph, Bayesian approaches to modeling the conditional dependence between multiple diagnostic tests, Biometrics 57 (2001) 208–217.